

This document is published at:

Nazir, S., Yousaf, M.H, Nebel, J.C. y Velastín, S.A. (2019). Dynamic Spatio-Temporal Bag of Expressions (D-STBoE) Model for Human Action Recognition. *Sensors*, 19(12), 2790.

DOI: <https://doi.org/10.3390/s19122790>



This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

## Article

# Dynamic Spatio-Temporal Bag of Expressions (D-STBoE) Model for Human Action Recognition

Saima Nazir <sup>1,2,3,4</sup> , Muhammad Haroon Yousaf <sup>5</sup> , Jean-Christophe Nebel <sup>6</sup>   
and Sergio A. Velastin <sup>3,4,7,\*</sup> 

<sup>1</sup> Department of Software Engineering, Fatima Jinnah Women University, Rawalpindi 46000, Pakistan; drsaima.nazir@fjwu.edu.pk

<sup>2</sup> Software Engineering Department, University of Engineering and Technology, Taxila 47050, Pakistan

<sup>3</sup> School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK

<sup>4</sup> Cortexica Vision Systems Ltd., London SE1 9LQ, UK

<sup>5</sup> Computer Engineering Department, University of Engineering and Technology, Taxila 47050, Pakistan; haroon.yousaf@uettaxila.edu.pk

<sup>6</sup> School of Computer Science and Mathematics, Kingston University, London KT1 2EE, UK; j.nebel@kingston.ac.uk

<sup>7</sup> Department of Computer Science, Universidad Carlos III de Madrid, Leganés, 28911 Madrid, Spain

\* Correspondence: sergio.velastin@ieee.org

Received: 16 May 2019; Accepted: 19 June 2019; Published: 21 June 2019



**Abstract:** Human action recognition (HAR) has emerged as a core research domain for video understanding and analysis, thus attracting many researchers. Although significant results have been achieved in simple scenarios, HAR is still a challenging task due to issues associated with view independence, occlusion and inter-class variation observed in realistic scenarios. In previous research efforts, the classical bag of visual words approach along with its variations has been widely used. In this paper, we propose a Dynamic Spatio-Temporal Bag of Expressions (D-STBoE) model for human action recognition without compromising the strengths of the classical bag of visual words approach. Expressions are formed based on the density of a spatio-temporal cube of a visual word. To handle inter-class variation, we use class-specific visual word representation for visual expression generation. In contrast to the Bag of Expressions (BoE) model, the formation of visual expressions is based on the density of spatio-temporal cubes built around each visual word, as constructing neighborhoods with a fixed number of neighbors could include non-relevant information making a visual expression less discriminative in scenarios with occlusion and changing viewpoints. Thus, the proposed approach makes the model more robust to occlusion and changing viewpoint challenges present in realistic scenarios. Furthermore, we train a multi-class Support Vector Machine (SVM) for classifying bag of expressions into action classes. Comprehensive experiments on four publicly available datasets: KTH, UCF Sports, UCF11 and UCF50 show that the proposed model outperforms existing state-of-the-art human action recognition methods in term of accuracy to 99.21%, 98.60%, 96.94 and 94.10%, respectively.

**Keywords:** human action recognition; Bag of Words (BoW); Bag of Expressions (BoE); spatio-temporal; dynamic neighborhood

## 1. Introduction

During the last decade, human action recognition (HAR) has become a well-developing research area and has received substantial attention. It has been applied in a wide range of applications such as sports video analysis, content-based video retrieval, video surveillance, scene recognition, video

annotation and human-computer interaction [1]. In particular, there is strong interest in the use of video and depth sensors to monitor the activity of elderly or disabled people in domestic environments to support their independence in a safe environment. The use of dedicated intelligent sensors in such cases could also contribute to maintaining privacy by avoiding the need to stream potentially intrusive video feeds to external human carers. The main aim of action recognition is to classify the action present in a test video irrespective of actors' locations. Action recognition in realistic scenarios is clearly much more challenging than in constrained environments because background objects can interfere and be mistaken as features of interest.

State-of-the-art approaches have shown notable increases in performance in realistic and complex scenarios. However, there are still some unsolved issues such as background clutter, viewpoint variation, illumination change and class variability [2]. Recently, significant progress has been demonstrated with spatio-temporal feature representation along with variations of the most popular and widely used bag of visual words approaches (BoVW) [3], which have the ability to handle viewpoint independence, occlusion and scale invariance [4,5]. Therefore, there has been a growing interest in exploring the potential of possible variants of the classical BoVW approach, which characterizes actions using a histogram of feature occurrence after clustering [6].

However, since BoVW approaches only consider the frequency of each visual word and ignore the spatio-temporal relation among visual words, such representation is not able to exploit the contextual relationship between visual words. To overcome this drawback, we propose an extension of our previously proposed Bag of Expressions (BoE) model [7], which incorporates contextual relationships of visual words while preserving inherent qualities of the classical BoVW approach. The main idea of visual expression formation is to store the spatio-temporal contextual information that is usually lost during formation of visual words by encoding neighboring spatio-temporal interest points' information.

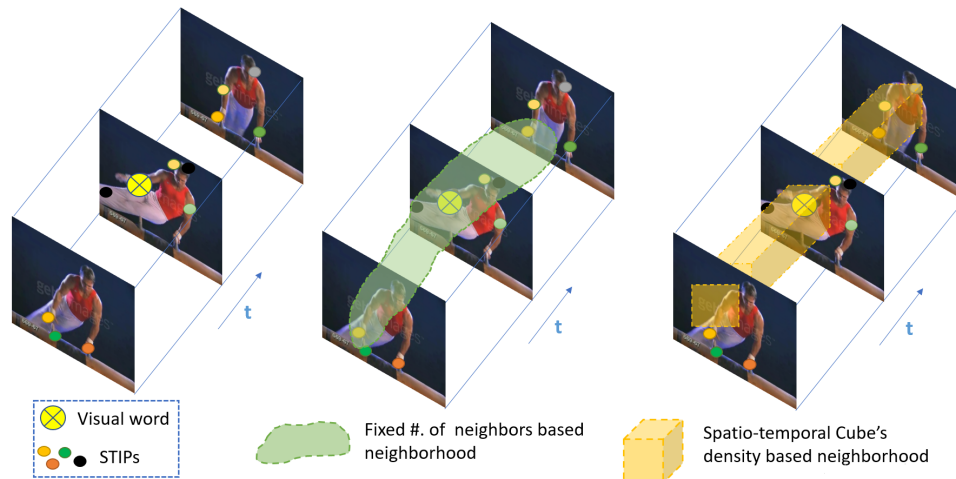
Both Kovashka et al. [8] and Gilbert et al. [9] have shown that performance can be improved by the inclusion of neighborhood amongst visual words. Although Reference [8] managed to preserve scale invariance, their solution led to a decline in the capability of handling different viewpoints and occlusions. Alternatively, the inclusion of neighborhoods by Gilbert et al. [9] affected the ability to handle scale and they had to integrate a hierarchy to preserve some scale invariance.

Constructing neighborhoods with a specific number of neighbors, as in Reference [7], can reduce performance in scenarios with occlusion and changing background. For example, under partial occlusion, neighborhoods would contain irrelevant neighbors which would add noise, making them less discriminative. We propose to overcome this by dynamically estimating the number of suitable neighbors, as described in Section 3.4.1, according to the local density of neighboring Spatio-Temporal Interest Points (STIPs) in a Spatio-Temporal Cube (ST-Cube). Figure 1 illustrates the difference between a neighborhood-based approach that uses a fixed number of neighbors with one that uses a spatio-temporal cube's density. In case of an viewpoint variation, the likelihood of obtaining the same visual expression representation is higher with an ST-Cube based neighborhood.

Kovashka et al. [8] were unable to address viewpoint changes because of restrictions they placed in the formation of neighborhoods. To overcome this, we propose to form a visual expression as an independent pair of visual words and neighboring STIPs. Using the density of a spatio-temporal cube of size  $n$ , each neighboring STIP in the ST-cube is paired with a visual word to obtain  $n$  different visual expressions. This provides a local representation of visual expressions and enables some tolerance to viewpoint variation and occlusion.

To support view independence, we represent each action as a visual expression and, as in Reference [7], only the frequencies of these visual expressions are stored. They are formed in such a way that they represent the distribution of visual expressions in the spatio-temporal domain. In contrast to Reference [7], visual expressions are constructed to incorporate the spatio-temporal contextual information of the visual words by combining the visual words and the neighboring STIPs present in each spatio-temporal cube. Similarly, these visual expression representations discard all information

related to other visual words and only consider the relationship between a visual word and its neighboring STIPs in the ST-Cube. It focuses on the individual contribution of visual expression and is expected to enhance discriminative power on occlusion. As will be shown here, the proposed model outperforms state-of-the-art results for KTH, UCF Sports, UCF11 and UCF50 datasets.



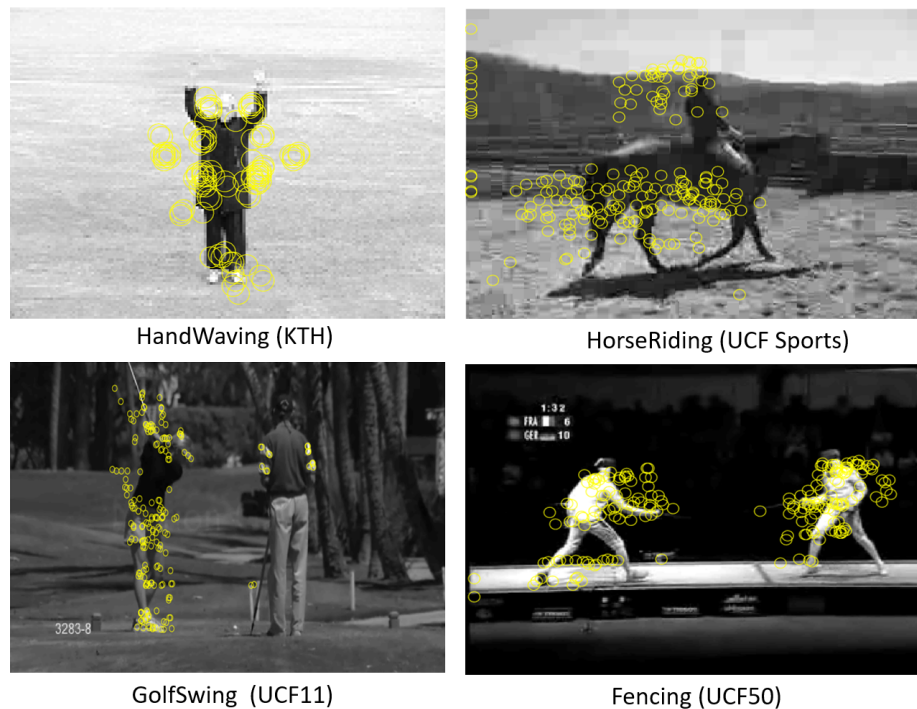
**Figure 1.** Illustration of STIPs (**left**), neighborhoods based on a fixed number of neighbors (**middle**) and a spatio-temporal cube's density (**right**). For a fixed number of neighbors, each visual word neighborhood is formed based on the same number of STIPs falling in that neighborhood. When using spatio-temporal cube's density, the neighborhood is formed based on the number of STIPs falling in the spatio-temporal cube (ST-Cube).

The rest of the paper is organized as follows. We discuss related work in Section 2. Section 3 describes in detail the proposed dynamic spatio-temporal Bag of Expressions (D-STBoE) model for human action recognition. Evaluation of parameters, experimentation results and discussion are provided in Section 4, followed by conclusions and suggestions for future work in Section 5.

## 2. Related Work

Bag of visual words (BoVW) along with its variations has become a popular approach for many visual interpretation tasks, including human action recognition. In BoVW, video representation is obtained using three steps: feature representation, dictionary learning and feature encoding. In such representation, performance is highly dependent on the discriminative power of each step. As work on action recognition using a BoVW approach is too broad to be covered here, we focus on the related approaches for feature representation and neighborhood features.

Feature representation for video, including holistic and local representation approaches, have been reviewed in detail in survey papers such as Reference [1]. Holistic representations use a global representation of human body, structure, shape and movement [10–12], while local representations are based on the extraction of local features [13]. Local representation is favored nowadays because holistic ones are too rigid to capture viewpoint variation, illumination changes and occlusions, which is typical of realistic scenarios [14,15]. Local feature representation emerges as an outcome of the spatio-temporal interest point detector proposed by Laptev et al. [16]. Other traditional feature detectors like Harris corner detector [17], Gabor filtering [18] and the scale invariant feature transform (SIFT) detector [19] are also popular interest point detectors. Based on the popular Harris corner detector, 3D Harris spatio-temporal interest point detector was proposed by Laptev et al. [16] to capture significant variations in both space and time, which has made it particularly applicable to action recognition. An example of such interest points is shown in Figure 2.



**Figure 2.** Detected Spatio-temporal interest points (shown in the space domain only, for clarity). Examples from KTH [16,20–23], UCF-Sports [24–26], UCF-11 [27,28] and UCF-50 [29,30] datasets.

Local descriptors are then computed for these detected interest points. Local interest point descriptors can rely on 3D cuboid or trajectories. 3D HoG (Histogram of Gradients) [31], inspired by HoG [32], is used to integrate motion information in the spatio-temporal domain. Histogram of optical flow (HoF) [33] encodes optical flow of pixel value motion in spatio-temporal local regions. Another popular 3D interest point descriptor is 3D SIFT [34].

After computing features, these are further transformed into visual words in a BoVW approach. Compared to text that is usually semi-structured, visual appearance is less structured and exhibits large variations when the same underlying visual pattern is present in varying illumination, different viewpoints and occluded environments. A possible solution to resolve the ambiguity of visual pattern representation is to consider its spatial-temporal context [7].

Yuan et al. [35] introduced the concept of visual phrase to overcome the underlying over-representation and under-representation challenges in visual word representation. Visual phrases are used to represent the spatial co-occurrence of visual words. Meng et al. [36] have used spatio-temporal cues for object search in video and find the top  $k$  trajectories that are the possible candidates containing the object.

For video representation, Zhao et al. [37] used deep spatial net (VGCNet [38]) and temporal net (Two-Stream ConNets [39]). They extracted the proposed frame-diff layer and convolutional layers and pooled them with trajectory temporal pooling and line temporal pooling strategies. These descriptors were further pooled with VLAD (Vector of Locally Aggregated Descriptors) for video representation. Xu et al. [40] combined a trainable VLAD encoding process and the RCNs (randomly connected neurons) architecture and developed a novel SeqVLAD (Sequential VLAD). Murtaza et al. [41] also proposed a novel encoding algorithm, DA-VLAD, for video representation by exploiting the code-word representation of each action.

Rahmani et al. [42] proposed a novel deep fully-connected neural network for the recognition of human actions from novel viewpoints which they called non-linear knowledge transfer model (R-NKTM). By using dense trajectories for feature representation, their proposed model performed better than traditional methods. In addition, they stated that performance could be further improved by combining multiple feature representation approaches like dense trajectories, HOG, HOF and motion boundary histogram.



Wang et al. [43] trained a deep neural network using spatio-temporal features. They used a Convolution Neural Network (CNN) and Long Short Term Memory (LSTM) units for the extraction of spatial and temporal motion information respectively. Local and semantics characteristics were used to distinguish an object from its background and in addition its spatial information was extracted using a CNN. Different spatial and temporal positions in video were learned using a temporal-wise attention model.

Khan et al. [44] improved the performance of human action recognition methods using a hybrid feature extraction model. For action classification, they used a multilayer neural network on selected fused features and achieved good performance on the KTH dataset.

Many researchers have extended the traditional 2D convolutional neural network to integrate temporal information. 3D CNN was proposed by Hara et al. [45] to include temporal information using a 3D kernel to extract information from both the spatial and the temporal domains. Limiting spatial space looks reasonable whereas, for the extraction of motion information, limiting time domain to a few frames seems deficient. Ng et al. [46] proposed that instead of learning temporal information from few frames, temporal pooling can be applied to integrate the temporal domain information in network layers.

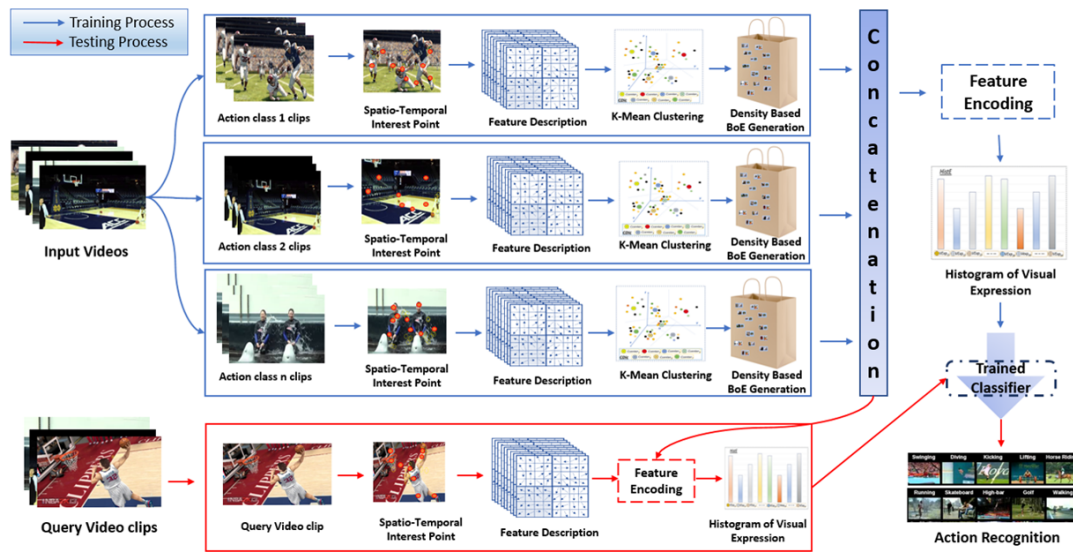
Simonyan et al. [39] extended the convolutional neural network by proposing a two-stream architecture. Similar to the way human vision cortex learns appearance and motion information using ventral (object recognition) and dorsal (motion recognition) streams, they used two pathways to learn spatial and temporal information for appearance and motion respectively. The spatial stream was learned on a model pretrained on ImageNet while the temporal stream was trained on an action recognition dataset to address the unavailability of pretrained models on flow information.

Following the success of the two-stream architecture, Feichtenhofer et al. [47] extended the two-stream CNN and proposed a spatio-temporal residual network. Motivated by the success of ResNet in ILSVRC 2015, they pretrained ResNet on ImageNet for action recognition. They embedded the residual connection between appearance and motion streams to allow the hierarchical learning of spatiotemporal features for both streams.

Hara et al. [45] also utilized the ResNet model for human action classification. They made use of 3D convolutional kernels to integrate temporal information in a residual network. They trained and evaluated their proposed 3D ResNet model on relatively large action recognition datasets (Kinetics and Activity Net). Temporal ResNet [48] also integrated temporal information in a spatial residual network by modifying the basic residual function using an additional temporal conv block.

### 3. Dynamic Spatio-Temporal Bag of Expressions (D-STBoE) Model

In this section, we present a dynamic spatio-temporal Bag of Expressions (D-STBoE) model for human action recognition. The proposed model recognizes human actions performed in complex scenarios by following the pipeline shown in Figure 3. This figure is divided into two sections, that is, training and testing processes. For training, first each video is described using a local feature representation, that is, spatio-temporal interest point's descriptions. We note that not only is the method independent from the type of features used but it can also be used with features extracted from pre-trained deep networks. For simplicity and to allow comparison with competing methods, here we have used well-established features. Second, for each action class, the widely used k-means clustering algorithm is employed to produce visual words from which a spatio-temporal cube's density based Bag of Expressions is generated. Third, a histogram of visual expressions is constructed for each class by encoding the training videos' feature representation using the concatenated visual expression dictionary. Finally, those histograms are used to train a supervised classifier. The testing process is similar: the video of interest is classified by first encoding its feature representation using the visual expression dictionary formed during the training process and then by submitting it to the action classifier. These steps are explained in more detail in the following sub-sections.



**Figure 3.** Dynamic Spatio-temporal Bag of Expressions (D-STBoE) Model for human action recognition.

### 3.1. Spatio-Temporal Interest Points Detection

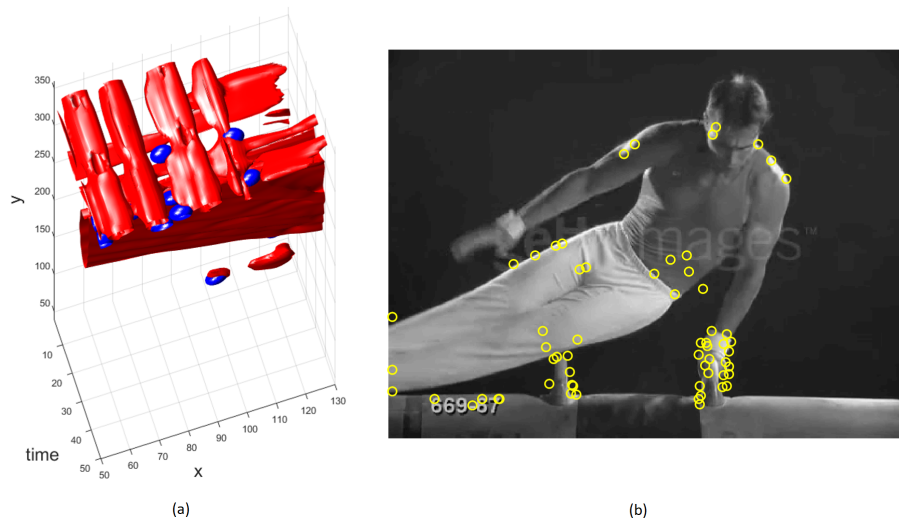
The feature detector aims to detect the distinctive locations of interest points which are used as a compact representation of the raw action video. Since in a video spatio-temporal interest points are considered to be both compact and salient for feature representation [14], 3D Harris is selected to detect interest points associated with regions with high-intensity variation in both space and time dimensions. More specifically, we take advantage of the modified Harris corner function proposed by Laptev et al. [16]:

$$D = \det \mu - m \text{trace}^3(\mu) \quad (1)$$

where  $m$  is a tunable sensitivity parameter and  $\mu$  is a  $3 \times 3$  spatio-temporal moment matrix:

$$\mu = G(:, \sigma^2, \tau^2) * \begin{Bmatrix} L_x^2 & L_x L_y & L_x L_t \\ L_x L_y & L_y^2 & L_y L_t \\ L_x L_t & L_y L_t & L_t^2 \end{Bmatrix} \quad (2)$$

with a spatial scale  $\sigma$  and temporal scale  $\tau$ . Here  $L$  is the first order spatio-temporal derivative and  $G$  is a Gaussian weighting function. As illustrated in Figure 4a where 3D plots show detected interest points for the detected STIPs (blue dots), those points are well localized in both space and time domains. For simplicity detected STIPs are shown in the space domain only in Figure 4b. This helps to obtain a rich representation for each video and enables to differentiate between other action classes and noise. For each video  $V$ , the set of extracted spatio-temporal interest points are given by  $p = \{p_i | p_i \in (x, y, t)_{i=1}^n\}$ , where  $n$  is the total number of interest points detected for a video  $V$ . Clearly, the value of  $n$  varies for each video depending on the number of interest points detected by the 3D Harris corner detector.



**Figure 4.** Spatio-temporal interest points for the Swing-Bench action from the UCF Sports dataset [24] (a) 3D Plot of detected interest points for Swing-Bench action (b) Pictorial representation of detected interest points for Swing-Bench action, shown only in the space domain for simplicity.

### 3.2. STIP Description using 3D SIFT

Feature descriptors are used to describe the detected interest points. Since 3-Dimensional Scale Invariant Feature Transform (3D SIFT) [34], an extension of SIFT [49] in the time domain, is able to represent points such that their representation is invariant in terms of scale, rotation, appearance and occlusion [50], we used that descriptor to characterize the points detected in the previous step. Scovanner et al. [34] obtained the 3D SIFT representation by computing the overall orientation of the neighborhood. In 3D, orientation and magnitude are calculated as:

$$Mag(x, y, t) = \sqrt{(L_x^2 + L_y^2 + L_t^2)} \quad (3)$$

$$\theta(x, y, t) = \tan^{-1} \left( \frac{L_y}{L_x} \right) \quad (4)$$

$$\phi(x, y, t) = \tan^{-1} \left( \frac{L_t}{\sqrt{L_x^2 + L_y^2}} \right) \quad (5)$$

$\phi(x, y, t)$  encodes the angle away from the 2D gradient direction and  $\theta(x, y, t)$  represents the angle in the 2D gradient. Once this is computed, the next step is to obtain the 3D SIFT descriptor representation by encoding orientations using sub-histograms. For each video  $V$ , the feature representations are denoted as  $FP = \{f_i, p_i | p_i \in (x, y, t)_{i=1}^n\}$  where  $f_i$  is the feature description for the spatio-temporal interest point  $p_i$ .

### 3.3. Class Specific Visual Word Dictionary

K-means clustering is a commonly used algorithm to automatically partition a dataset into  $k$  groups [1,51]. Let  $FP = \{F_1P_1, F_2P_2, F_3P_3, \dots, F_{tm}P_{tm}\}$  represent the obtained features where  $tm$  is the total number of videos for an action class,  $FP$  is a set of feature representation for the videos of an action class. We apply k-means clustering to obtain class-specific visual word representations, which correspond to the  $k$  cluster centers. They become the words of the class-specific dictionary,  $CDic$ .



### 3.4. Bag of Expressions Generation

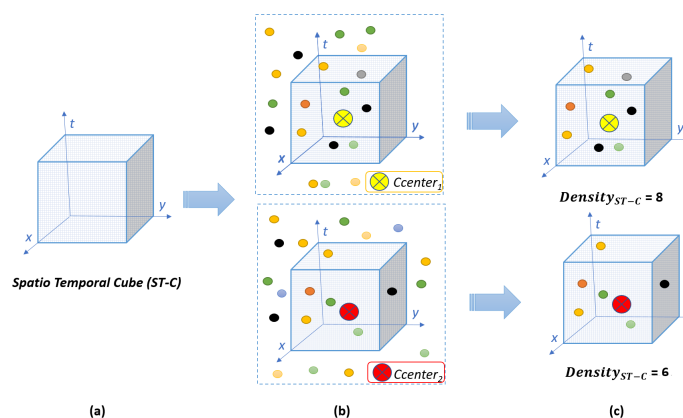
After the generation of class-specific visual *word* dictionaries, the next step is to generate a concatenated *visual expression* dictionary containing  $C$  class-specific visual expression dictionaries, where  $C$  is the total number of unique action classes.

Constructing a neighborhood with visual words with a fixed number of neighbors [7] affects performance in scenarios with occlusions and changing viewpoints since, for example, in the case of partial occlusion the same neighborhood could contain non-relevant (different class) neighbors, making the neighborhood less discriminative. We propose to set the number of suitable neighbors dynamically, as described in Section 3.4.1, according to the density of a spatio-temporal cube in the frame of interest. However, as such strategy may reduce some viewpoint independence, since it is less probable to obtain neighborhoods containing the same words, this is addressed by describing the neighborhood as an independent pair of neighbors. Each element in a pair contains an independent feature representation with respect to the other paired elements. Each visual word is paired with its neighboring interest points to form visual expressions. Such representation discards all information related to other words and only considers the relation between respective visual word and its neighboring STIP in a pair. Hence, this allows focusing on the individual contribution of a visual expression which should enhance its discrimination power when dealing with occlusions.

A bag of visual expressions is generated by incorporating a visual word's neighborhood information present in the spatio-temporal cube, as described in the following sub-sections. After the construction of a spatio-temporal cube around a visual word and the calculation of the density of the respective ST-Cube, Section 3.4.1, Section 3.4.2 describes the formation of visual expressions by pairing visual words with their respective neighboring STIPs in the ST-Cube. Finally, Section 3.4.3 explains the formation of a visual expression dictionary by using the  $C$  class-specific visual expression dictionary.

#### 3.4.1. Spatio-Temporal Cube

For each cluster center, ' $C_{center}$ ' (visual word), in a dictionary  $CDic$ , a spatio-temporal cube (ST-C) is formed containing the spatio-temporal interest points around the cluster centers. As shown in Figure 5a, the spatio-temporal cube is aligned with the 3D plane axes. This ST-Cube is constructed to define the neighborhood of the respective visual word to dynamically obtain a suitable neighborhood for each visual word, making it more discriminative. This spatio-temporal cube contains neighboring spatio-temporal interest points from the set  $P$ , where  $P$  is the set of interest points detected for an action class (as mentioned in Section 3.1). As shown in Figure 5b, the number of neighbors in a cube can vary and is also dependent on the size of the cube, denoted as  $CubeSize$ .



**Figure 5.** (a) Spatio-temporal Cube aligned with 3D plan axes (b) Neighboring spatio-temporal interest point around cluster centers (c) visual word density calculation.

The density of each visual word (also referred as cluster center) is calculated in the spatio-temporal cube where ST-C density is defined as:

$$Density_{ST-C} = \sum_j DenseX(p_j, Ccenter) \quad (6)$$

where  $DenseX = 1$  if  $p_j$  falls in the spatio-temporal cube of size  $CubeSize^3$ , otherwise  $DenseX = 0$ . In other words, visual word density is the number of spatio-temporal neighborhood interest points present in the spatio-temporal cube. Figure 5c illustrates the calculation of visual words' densities.

### 3.4.2. Visual Expression Formation

Visual expressions are formed to incorporate the contextual spatio-temporal neighboring information of a visual word, by combining the visual words and their respective neighboring spatio-temporal interest points present in the spatio-temporal cube. Let  $p_q$  be the neighboring spatio-temporal interest point of the visual word  $Ccenter_1$ . The associated visual expression is formed as:

$$VExp_{1q}(f, p) = [Ccenter_1(f_{Ccenter_1}, p_{Ccenter_1}), fp_q(f_q, p_q)] \quad (7)$$

The number of visual expressions formed against each visual word depends on its  $Density_{ST-C}$ . For example, as shown in Figure 6, if  $Ccenter_1$  has  $Density_{ST-C} = 8$ , then the number of visual expressions formed for  $Ccenter_1$  will be eight and are represented as  $\{VExp_{11}(f, p), VExp_{12}(f, p), VExp_{13}(f, p), \dots, VExp_{18}(f, p)\}$ .

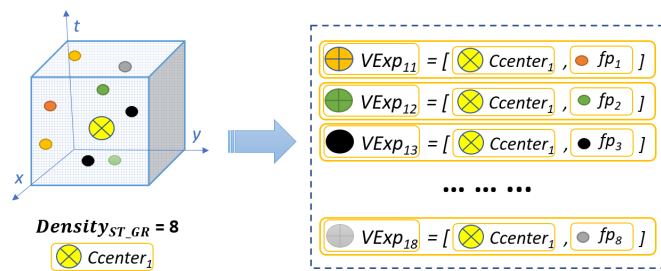
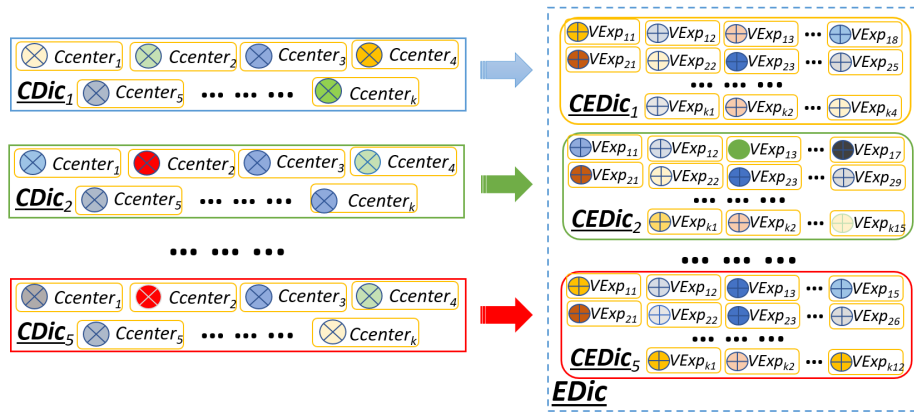


Figure 6. Visual expression formation for a visual word  $Ccenter_1$  of density 8.

### 3.4.3. Visual Expression Dictionary

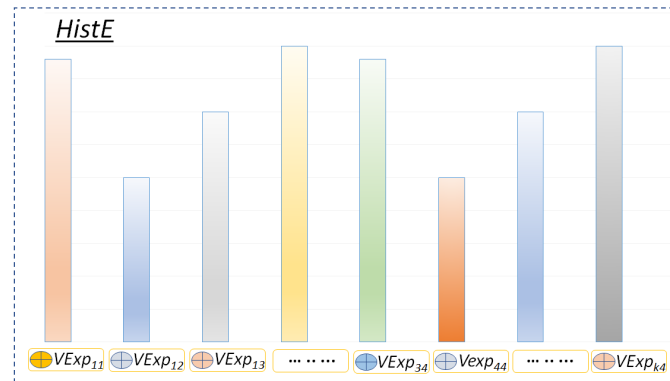
Similarly to the approach used in Reference [7], after obtaining visual expression representations for each visual word, they are concatenated to form a visual expression dictionary  $CEDic$ . In contrast to Reference [7], the size of  $CEDic$  depends on the number of visual words, that is,  $k$ , obtained for each action class, the number of classes and spatio-temporal cube's densities. As shown in Figure 7, assuming there are five action classes in an action recognition dataset, after applying k-means clustering (as discussed in Section 3.3), there will be five class specific dictionaries with  $k$  visual words for each  $CEDic$ . Then, the class-specific visual expression dictionaries,  $CEDics$ , are generated for each action class. Note that the size of each  $CEDic$  varies as it is dependent on the size of each visual word's ST-C density. Finally, these five  $CEDic$  are concatenated to obtain a visual expression dictionary, which is represented as  $EDic$ .



**Figure 7.** An example of visual expression dictionary (*EDic*) generation from class specific dictionaries (*CDics*).

### 3.5. Histogram of Visual Expressions Encoding

The proposed model uses histogram encoding [52] for assigning the local feature representation to the corresponding visual expression. Given a set of feature representations  $fp = f_i p_i |_{i=1}^n$  of a video  $V$ , let  $histE_s$  be the assignment of each feature representation  $f_s p_s$  to the corresponding visual expression  $VExp_r$  such that  $histE_s = \operatorname{argmin} \|f_s p_s - VExp_r\|^2$ , see Figure 8. The histogram encoding vector  $HistE \in R^r$  is such that  $[HistE]_r = |\{s : histE_s = r\}|$  where  $r$  is the corresponding  $VExp$  in *EDic*.



**Figure 8.** Histogram of visual expression representation.

### 3.6. Action Recognition

Since we use a multi-class support vector machine model for action recognition, we need to select both a kernel function and a multi-class model. We evaluated multi-class SVM models using three different kernel functions, that is, Gaussian kernel, linear kernel and polynomial kernel, which are defined, respectively, as:

$$GK(x, y) = \exp(-\|x - y\|^2) \quad (8)$$

$$LK(x, y) = x' y \quad (9)$$

$$PK(x, y) = (1 + xy)^2 \quad (10)$$

Multi-class SVMs models use different types of coding design schemes to learn [53]. A video of interest is assigned the action class label  $\hat{C}$  that maximizes aggregation of the losses for the  $L$  binary learner and defined as:

$$\hat{C} = \operatorname{argmin}_c \sum_{l=1}^L |M_{cl}| g(M_{cl}, S_l) \quad (11)$$

where  $M$  is the coding design scheme matrix with elements  $M_{cl}$ ,  $S_l$  and is the predicted classification score for the positive class of learner  $l$  and  $C$  is the total number of unique action classes. Six different coding design metrics were considered for learning a binary learner:

1. **One vs. one:** The multiclass model learns  $C(C - 1)$  binary learners. For this coding design matrix, one class is positive, another is negative and the rest is ignored.
2. **One vs. All:** For each binary learner of a total  $C$  binary learners, one class is positive and the rest is negative.
3. **Ordinal:** Matrix elements are assigned such that for the first binary learner the first class is positive and others are negative. For the second binary learner, the first two classes are positive and the rest are negative and so on. It has  $C - 1$  binary learners.
4. **Binary Complete:** The multi-class SVMs model learns  $2^{C-1} - 1$  binary learners such that it assigns 1 or  $-1$  to each class provided that at least one class is assigned positive and one class is assigned negative.
5. **Dense Random:** Each element of the design matrix is assigned equal probability and has approximately  $10\log_2 C$  binary learners.
6. **Sparse Random:** It assigns 1 or  $-1$  to the class with 0.25 probability and 0 to class with 0.5 probability. It has approximately  $15\log_2 C$  binary learners.

The dynamic spatio-temporal based Bag of Expressions (D-STBoE) method is summarized in Algorithm 1.

---

**Algorithm 1** D-STBoE algorithm for Human Action Recognition

---

**Require:** Training video dataset  $VS_{train} = \{VS_1, VS_2, VS_3, \dots, VS_C\}$  where  $C$  is the number of action classes,

Training labels  $Label_{train}$ ,  
Testing video dataset  $VS_{test}$

```

1:  $EDic \leftarrow []$ 
2: for  $z = 1:C$  do
3:   for each video in  $VS_z$  do
4:     obtain feature representation  $FP$ 
5:     do k-mean clustering to obtain  $CDic_z$ 
6:     for each visual word  $Ccenter_w$  in  $CDic_z$  do
7:       for  $u = 1 : Density_{Ccenter_w}$  do
8:         compute  $VExp_{uw}$ 
9:          $CEDic_z = concatenate[CDic_z, VExp_{uw}]$ 
10:     $EDic = concatenate[EDic, CEDic_z]$ 
11: for each video  $v$  in training set  $VS_{train}$  do
12:   feature encoding using  $EDic$  to get  $HistE_{train}$ 
13:  $TrainC = TrainClassifier(HistE_{train}, Label_{train})$ 
14: for each video  $V_{test}$  in training set  $VS_{test}$  do
15:   obtain feature representation  $FP$ 
16:   feature encoding using  $EDic$  to get  $HistE_{test}$ 
17:    $Label_{test} = Predict(TrainC, HistE_{test})$ 
return  $Label_{test}$ 

```

---

## 4. Experimentation Results and Discussion

### 4.1. Datasets

The proposed model was evaluated on four benchmark datasets for human action recognition, that is, KTH, UCF Sport, UCF11 and UCF50. Although KTH [21] contains actions performed in simple scenarios and as a consequence it is not considered as challenging by state-of-the-art human action recognition methods, it is included for comparison with related work. It consists of videos involving a

single actor performing single actions. It contains 6 different types of actions. They are recorded in a constrained environment with a static background. The dataset is divided into training and testing videos based on actor identity [21]. Videos involving 16 different actors are used for training, while the remaining, involving 9 different actors, are used for testing. Performance is reported using average accuracy over all classes.

The **UCF Sports** [25] dataset comprises 150 videos taken from actual sports broadcasts. It contains 10 different types of actions. It is a challenging dataset since it is captured in realistic environments which include wide ranges of views and scenes. Since each action is performed by different actors, each individual action is performed in many different manners. We used the leave-one-sample-out validation scheme: one original video is used for testing and the other 149 videos for training. Performance is evaluated according to average accuracy, that is, accuracy over all classes.

**UCF11** [28], previously known as the YouTube dataset, is a very challenging dataset as it contains videos taken from the web. They are captured in unconstrained environments and include challenges such as large variation of illumination changes, a mix of steady and shaking cameras, cluttered backgrounds, low resolution and variation of object scale. It comprises 11 types of actions. Each action class has 25 groups, each containing at least 4 video clips. Each group's video clips share features such as having been captured in a similar environment from the same viewpoint and having been performed by the same actor. Similar to Reference [25], we used the leave-one-group-out validation method and performance is reported using average accuracy over all classes.

**UCF50** is also a very challenging dataset captured in realistic environments with large variations in viewpoints, backgrounds, camera motions, object appearances and poses. Similarly to UCF11, for each action class in UCF50, video clips are grouped into 25 groups each containing at least 4 video clips. Each group shares some features, like similar environment, same actor and similar viewpoints. As proposed in References [25,30], a Leave One Group Out evaluation method is used for evaluation and performance is reported using average accuracy over all classes.

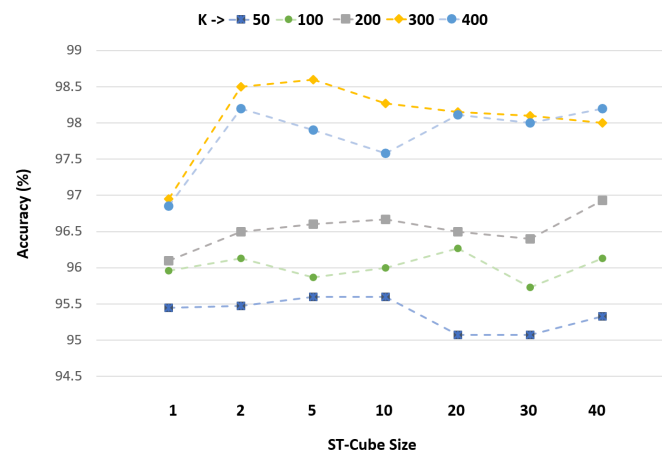
#### 4.2. Feature Extraction and Parameter Tuning

We have extracted spatio-temporal interest points using the 3D Harris STIP detector with standard values, that is,  $\sigma = 4$ ,  $\tau = 2$  and  $m = 0.001$  [15]. These points are described by 3D SIFT with 2056 dimensions. We optimized parameters involved in the proposed model on UCF Sports dataset as it is a challenging dataset containing a wide range of actions and scenes. We performed a set of experiments to choose the best  $k$  visual words representation for the class-specific dictionary.

We also evaluated the proposed D-STBoE model performance by varying ST-Cube size. Figure 9 shows that the increase in class-specific dictionary (CDic) size (i.e., value of  $k$ ) results in improved performance irrespectively of ST-Cube size. It also reveals that an increase in ST-Cube size tends to decrease performance after ST-Cube size = 5. Indeed, a larger size is likely to increase the number of non-relevant neighbors contributing to the visual expression formation, hence making it less discriminant. However, a smaller ST-Cube size may fail to integrate relevant neighbors, hence making the visual expression incomplete. Performance is even more dependent on the ' $k$ ' parameter as an insufficient dictionary size leads to poorer accuracy. As our experiments show that up to  $k = 300$  there is significant performance improvement, that value was selected. Similarly, an ST-Cube Size of 5 was chosen to produce optimal performance.

To analyze the computational complexity of the proposed method, we have evaluated the impact of the parameter that has the most significant contribution in terms of computational time. Following measurements of the average running time for the generation of a class-specific dictionary as described in Section 3.3, Table 1 shows that an increased number of cluster's centers ( $k$ ) for k-means clustering leads to increased processing time.





**Figure 9.**  $k$  and  $ST$ -Cube size parameter evaluation for proposed D-STBoE model on the UCF Sports dataset.

**Table 1.** Computation time with respect to the class specific dictionary size ( $k$ ).

Size (CDic)	Time (s)
K = 50	92.20
K = 100	198.35
K = 200	369.45
K = 300	524.12

In the next set of experiments, we evaluated the performance of different multi-class SVM models used for action recognition by testing three different kernel functions and six coding design matrices for the binary learner, as outlined in Section 3.6. As expected from Reference [54], results show that the Gaussian kernel outperforms linear and polynomial kernels, see Table 2. Moreover, the One versus All coding design matrix shows better performance than the other matrices.

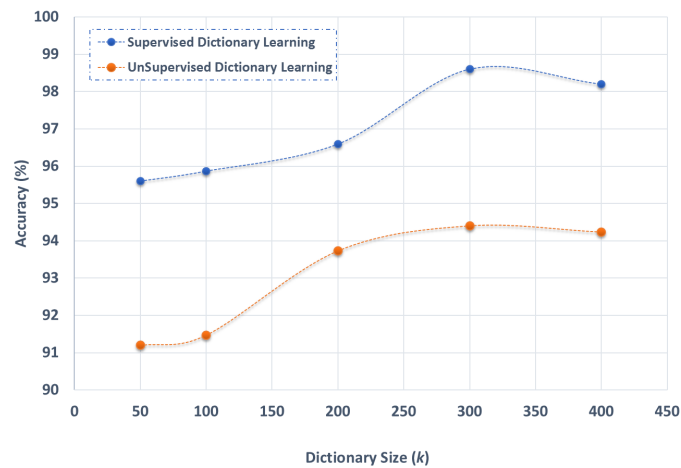
**Table 2.** Evaluation of Multi-class SVM model for action recognition on the UCF Sports dataset.

Coding Design Matrix/ Kernel Function	Gaussian Kernel	Linear Kernel	Polynomial Kernel
One vs. One	96.7	95.5	94.5
One vs. All	98.6	97.1	96.7
Ordinal	95.7	94.7	94.0
Binary Complete	97.7	96.4	95.8
Dense Random	98.5	97.2	96.6
Sparse Random	98.5	97.7	96.7

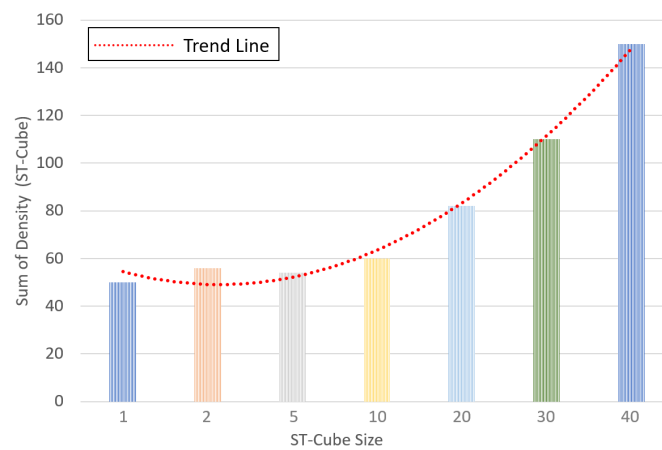
#### 4.3. Contribution of Each Stage

This section considers the contribution of each stage in the proposed model, as tested on the UCF Sports dataset. As presented in Section 3.3, we used a supervised dictionary learning approach (using  $k$ -means clustering) to learn the class-specific visual word representation for each class. Here, we evaluate the impact of learning the ' $C$ ' class specific dictionaries versus one unsupervised dictionary. Figure 10 shows that a significant improvement is brought by using a supervised dictionary learning approach in comparison with the unsupervised one. Although an increase in dictionary size leads in both cases to improved performance, the supervised approach outperforms the unsupervised one consistently.

The trend line in Figure 11 shows that the density of spatio-temporal varies approximately quadratically with  $ST$ -Cube size. Therefore, we can conclude that an increase in cube size will also result in an increase in visual expression dictionary size.

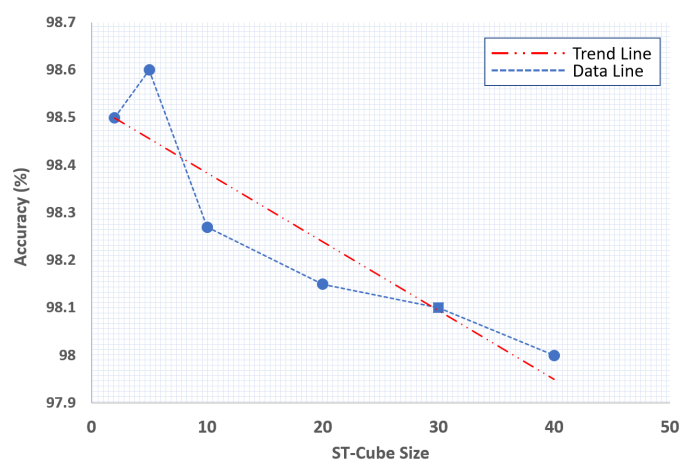


**Figure 10.** Evaluation of dictionary learning approach on the UCF Sports dataset.



**Figure 11.** Relationship between ST-Cube size and sum of ST-Cube Density.

However, this trend does not result in increased accuracy, as illustrated by Figure 12. If we compare this trend with the effect of visual word dictionary size on performance (refer to Figure 10), we can conclude that there is no relationship between visual expression dictionary size and visual word dictionary size in terms of achieved performance.



**Figure 12.** Spatio-temporal cube size evaluation.

#### 4.4. Comparison with the State-of-the-Art

The experiment section concludes with a comparison of the proposed model with state-of-the-art methods, including recent deep learning based methods. It should be noted that D-STBoE parameters were optimized for UCF Sports dataset as discussed in Section 3.4.2.

##### 4.4.1. Evaluation on the KTH Dataset

Table 3 shows that, in terms of average accuracy, D-STBoE and our previously proposed BoE perform better than all state-of-the-art methods with the exception of the Multilayer neural network, which achieved a slightly better performance of 99.80% against 99.21% for D-STBoE.

**Table 3.** Accuracy evaluation of our method compared to state-of-the-art methods on the KTH Dataset.

Author	Method	Results
Proposed	Dynamic Spatio-temporal Bag of Expressions (D-STBoE) Model	99.21
[7]	Bag of Expression (BoE)	99.51
[44]	Multilayer neural network	<b>99.80</b>
[55]	Foreground Trajectory extraction method	97.50
[56]	Power difference template representation	95.18
[57]	Direction-dependent feature pairs and non-negative low-rank sparse model	94.99
[58]	Space-time robust representation	94.60

##### 4.4.2. Evaluation on the UCF Sports Dataset

As shown in Table 4, D-STBoE outperforms all state-of-the-art methods, achieving an additional 2.6% accuracy when compared with the second best performer. This success should, however, be tempered by the fact that D-STBoE was optimised for this dataset. It should also be noted that two other 'traditional' methods outperform the deep learning approaches on this dataset.

**Table 4.** Accuracy evaluation of our method compared to state-of-the-art methods on the UCF Sports Dataset.

Author	Method	Results
Proposed	Dynamic Spatio-temporal Bag of Expressions (D-STBoE) Model	<b>98.60</b>
[7]	Bag of Expression (BoE)	97.33
[43]	Spatio-temporal features with deep neural network	91.80
[42]	Robust non-linear knowledge transfer model (R-NKTM)	90.00
[59]	Universal multi-view dictionary	91.00
[55]	Foreground Trajectory extraction method	91.40
[60]	Holistic and motion cues based approach	95.60
[61]	3D-TCCHOGAC and 3D-HOOFGAC feature descriptors	96.00
[56]	Power difference template representation	88.60
[62]	Factored matrices of dual tensors	92.70
[63]	Multi-Region two stream R-CNN	95.50
[64]	Spatio-temporal tube of maximum mutual information	90.70
[65]	Local motion and group sparsity-based approach	89.70
[66]	Dense trajectories and motion boundary descriptors	88.00
[67]	Sparse representation and supervised class specific dictionary learning	86.60
[68]	Invariant spatio-temporal features with independent subspace analysis	86.50
[69]	Person detector along with BOF approach	86.70
[8]	hierarchy of discriminative space-time neighborhood features	87.30

##### 4.4.3. Evaluation on the UCF11 and UCF50 Datasets

We compared the proposed model with the state of the art for the UCF11 and UCF50 datasets, with successful methods that have proven particularly suitable in dealing with the many practical problems represented in these realistic datasets. As shown in Table 5, our method outperforms

significantly all the state-of-the-art approaches for UCF11 dataset with the exception of a deep neural network, which delivers an accuracy 1.82% higher.

Table 6 shows the comparison of proposed approach for UCF50 dataset. Even in this very challenging and realistic dataset, the D-STBoE model shows reliable recognition performance. This further indicates that the class-specific visual expression formation is discriminative and that the D-STBoE model is effective for action recognition.

**Table 5.** Accuracy evaluation of our method compared to state-of-the-art methods on the UCF11 Dataset.

Author	Method	Results
Proposed	Dynamic Spatio-temporal Bag of Expressions (D-STBoE) Model	96.94
[43]	Spatio-temporal features with deep neural network	<b>98.76</b>
[59]	Universal multi-view dictionary	85.90
[55]	Foreground Trajectory extraction method	91.37
[70]	Graph-based multiple-instance learning	84.60
[65]	Local motion and group sparsity-based approach	86.10
[66]	Dense trajectories and motion boundary descriptors	84.10
[68]	Invariant spatio-temporal features with independent subspace analysis	75.80

**Table 6.** Accuracy evaluation of our method compared to state-of-the-art methods on the UCF50 Dataset.

Author	Method	Results
Proposed	Dynamic Spatio-temporal Bag of Expressions (D-STBoE) Model	<b>94.10</b>
[71]	HMG + iDT Descriptor	93.00
[72]	Bag of Words and Fusion Methods	92.30
[5]	Dense Trajectories	91.70
[66]	Dense Trajectories and motion boundary descriptor	91.20

A particular strength of the proposed model is that it shows some degree of tolerance to occlusion and viewpoint variation. Figures 13 and 14 show some positive examples of the ‘Basketball’ action (UCF11 dataset) captured from different viewpoints and ‘Spiking’ action (UCF11 dataset) in a non-occluded and occluded environment respectively. As the proposed D-STBoE model generates a visual expression based on cuboid neighborhood density, it only focuses on relevant local information which enhances its discrimination power.



**Figure 13.** Positive examples of ‘Basketball’ action (UCF11 dataset [27]) captured from different viewpoints.



**Figure 14.** Positive examples of ‘Spiking’ action (UCF11 dataset [27]) in non-occluded and occluded environment.

## 5. Conclusions and Future Work

In this paper, a dynamic spatio-temporal Bag of Expressions model for human action recognition has been introduced. Not only does it preserve the inherent qualities of the classical BoVW approach but it also acquires contextual information that would be lost with a conventional formation of visual words representation. We proposed a new Bag of Expressions generation method based on the density of spatio-temporal cube of visual words. The new model addresses the inter-class variation challenge observed in realistic scenarios by using a class-specific representation for visual expression generation. Viewpoint variation and occlusion challenges are also partially addressed by considering the spatio-temporal relation of visual words representation. The model was tested on four publicly available datasets that is, KTH, UCF Sports, UCF11 and UCF50. D-STBoE consistently outperformed all state-of-the-art ‘traditional’ methods while achieving competitive results when compared to deep learning approaches, which demonstrated best performance on two of the four datasets. In terms of practical deployment, D-STBoE could be a competitive solution compared to deep learning based systems, since, unlike them, it does not require a large dataset for training that is, D-STBoE is better suited for practical applications where availability of training data is usually limited.

Future work will explore the potential advantage of using spatio-temporal cube’s density based neighborhood information for deep learning features and methods and also investigate exploiting spatio-temporal neighborhood information for action detection in even more realistic and complex scenarios.

**Author Contributions:** Conceptualization, methodology, software, validation, investigation, resources, visualization, writing—original draft preparation, S.N.; Data curation, formal analysis, writing—review and editing, all authors; Supervision; project administration; funding acquisition, M.H.Y. and S.A.V.

**Funding:** Sergio A. Velastin is grateful for funding received from the Universidad Carlos III de Madrid, the European Union’s Seventh Framework Programme for research, technological development and demonstration under grant agreement N° 600371, el Ministerio de Economía, Industria y Competitividad (COFUND2013-51509) el Ministerio de Educación, Cultura y Deporte (CEI-15-17) and Banco Santander. Muhammad Haroon Yousaf received funding from the Higher Education Commission, Pakistan for Swarm Robotics Lab under the National Centre for Robotics and Automation (NCRA). The authors also acknowledge support from the Directorate of ASR&TD, University of Engineering and Technology Taxila, Pakistan.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript or in the decision to publish the results.

## References

1. Poppe, R. A survey on vision-based human action recognition. *Image Vis. Comput.* **2010**, *28*, 976–990. [[CrossRef](#)]
2. Xia, L.; Chen, C.C.; Aggarwal, J. View invariant human action recognition using histograms of 3d joints. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Providence, RI, USA, 16–21 June 2012; pp. 20–27.



3. Nazir, S.; Yousaf, M.H.; Velastin, S.A. Feature Similarity and Frequency-Based Weighted Visual Words Codebook Learning Scheme for Human Action Recognition. In *Pacific-Rim Symposium on Image and Video Technology*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 326–336.
4. Lopes, A.P.B.; Oliveira, R.S.; de Almeida, J.M.; de Araújo, A.A. Spatio-temporal frames in a bag-of-visual-features approach for human actions recognition. In *Proceedings of the 2009 XXII Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI)*, Rio de Janeiro, Brazil, 11–15 October 2009; pp. 315–321.
5. Wang, H.; Oneata, D.; Verbeek, J.; Schmid, C. A robust and efficient video representation for action recognition. *Int. J. Comput. Vis.* **2016**, *119*, 219–238. [[CrossRef](#)]
6. Shao, L.; Mattivi, R. Feature detector and descriptor evaluation in human action recognition. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, Xi'an, China, 5–7 July 2010; pp. 477–484.
7. Nazir, S.; Yousaf, M.H.; Nebel, J.C.; Velastin, S.A. A Bag of Expression framework for improved human action recognition. *Pattern Recognit. Lett.* **2018**, *103*, 39–45. [[CrossRef](#)]
8. Kovashka, A.; Grauman, K. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, CA, USA, 13–18 June 2010; pp. 2046–2053.
9. Gilbert, A.; Illingworth, J.; Bowden, R. Action recognition using mined hierarchical compound features. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 883–897. [[CrossRef](#)] [[PubMed](#)]
10. Bobick, A.F.; Davis, J.W. The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 257–267. [[CrossRef](#)]
11. Tian, Y.; Cao, L.; Liu, Z.; Zhang, Z. Hierarchical filtered motion for action recognition in crowded videos. *IEEE Trans. Syst. Man Cybern. Part C* **2012**, *42*, 313–323. [[CrossRef](#)]
12. Murtaza, F.; Yousaf, M.H.; Velastin, S.A. Multi-view Human Action Recognition Using Histograms of Oriented Gradients (HOG) Description of Motion History Images (MHIs). In *Proceedings of the 2015 13th International Conference on Frontiers of Information Technology (FIT)*, Islamabad, Pakistan, 14–16 December 2015; pp. 297–302.
13. Nazir, S.; Yousaf, M.H.; Velastin, S.A. Evaluating a bag-of-visual features approach using spatio-temporal features for action recognition. *Comput. Electr. Eng.* **2018**, *72*, 660–669. [[CrossRef](#)]
14. Bregonzio, M.; Gong, S.; Xiang, T. Recognising action as clouds of space-time interest points. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Miami Beach, FL, USA, 20–25 June 2009; pp. 1948–1955.
15. Nazir, S.; Yousaf, M.H.; Velastin, S.A. Inter and intra class correlation analysis (IICCA) for human action recognition in realistic scenarios. In *Proceedings of the 8th International Conference of Pattern Recognition Systems (ICPRS 2017)*, Madrid, Spain, 11–13 July 2017.
16. Laptev, I. On space-time interest points. *Int. J. Comput. Vis.* **2005**, *64*, 107–123. [[CrossRef](#)]
17. Harris, C.; Stephens, M. A combined corner and edge detector. In *Proceedings of the Alvey Vision Conference*, Manchester, UK, 31 August 1988; Volume 15, pp. 10–5244.
18. Dollár, P.; Rabaud, V.; Cottrell, G.; Belongie, S. Behavior recognition via sparse spatio-temporal features. In *Proceedings of the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, Beijing, China, 15–16 October 2005; pp. 65–72.
19. Lowe, D.G. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, Kerkyra, Greece, 20–27 September 1999; pp. 1150–1157.
20. Laptev, I.; Caputo, B. Recognition of Human Actions. 2005. Available online: <http://www.nada.kth.se/cvap/actions/> (accessed on 15 June 2019).
21. Schuldts, C.; Laptev, I.; Caputo, B. Recognizing human actions: A local SVM approach. In *Proceedings of the 17th International Conference on Pattern Recognition*, Cambridge, UK, 23–26 August 2004; pp. 32–36.
22. Laptev, I. Local Spatio-Temporal Image Features for Motion Interpretation. Ph.D. Thesis, Numerisk Analys Och Datalogi, KTH, Stockholm, Sweden 2004.
23. Laptev, I.; Lindeberg, T. Velocity adaptation of space-time interest points. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR)*, British Machine Vis Assoc, Cambridge, UK, 23–26 August 2004; pp. 52–56.
24. Rodriguez, M.D.; Ahmed, J.; Shah, M. UCF Sports Action Data Set. 2011. Available online: [https://www.crcv.ucf.edu/data/UCF\\_Sports\\_Action.php](https://www.crcv.ucf.edu/data/UCF_Sports_Action.php) (accessed on 15 June 2019).

25. Rodriguez, M.D.; Ahmed, J.; Shah, M. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
26. Soomro, K.; Zamir, A.R. Action recognition in realistic sports videos. In *Computer Vision in Sports*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 181–208.
27. Liu, J.; Shah, M. UCF YouTube Action Data Set. 2011. Available online: [https://www.crcv.ucf.edu/data/UCF\\_YouTube\\_Action.php](https://www.crcv.ucf.edu/data/UCF_YouTube_Action.php) (accessed on 15 June 2019).
28. Liu, J.; Luo, J.; Shah, M. Recognizing realistic actions from videos “in the wild”. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 1996–2003.
29. Reddy, K.; Shah, M. UCF50—Action Recognition Data Set. 2012. Available online: <https://www.crcv.ucf.edu/data/UCF50.php> (accessed on 15 June 2019).
30. Reddy, K.K.; Shah, M. Recognizing 50 human action categories of web videos. *Mach. Vis. Appl.* **2013**, *24*, 971–981. [[CrossRef](#)]
31. Klaser, A.; Marszałek, M.; Schmid, C. A spatio-temporal descriptor based on 3d-gradients. In Proceedings of the BMVC 2008-19th British Machine Vision Conference, British Machine Vision Association, Leeds, UK, 1–4 September 2008; pp. 1–275.
32. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; pp. 886–893.
33. Laptev, I.; Marszalek, M.; Schmid, C.; Rozenfeld, B. Learning realistic human actions from movies. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
34. Scovanner, P.; Ali, S.; Shah, M. A 3-dimensional sift descriptor and its application to action recognition. In Proceedings of the 15th ACM international conference on Multimedia, Augsburg, Germany, 25–29 September 2007; pp. 357–360.
35. Yuan, J.; Wu, Y.; Yang, M. Discovery of collocation patterns: From visual words to visual phrases. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
36. Meng, J.; Yuan, J.; Yang, J.; Wang, G.; Tan, Y.P. Object instance search in videos via spatio-temporal trajectory discovery. *IEEE Trans. Multimed.* **2016**, *18*, 116–127. [[CrossRef](#)]
37. Zhao, S.; Liu, Y.; Han, Y.; Hong, R.; Hu, Q.; Tian, Q. Pooling the convolutional layers in deep convnets for video action recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *28*, 1839–1849. [[CrossRef](#)]
38. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
39. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems 27 (NIPS Proceedings)*; MIT Press: Cambridge, MA, USA, 2014; pp. 568–576.
40. Xu, Y.; Han, Y.; Hong, R.; Tian, Q. Sequential video VLAD: Training the aggregation locally and temporally. *IEEE Trans. Image Process.* **2018**, *27*, 4933–4944. [[CrossRef](#)] [[PubMed](#)]
41. Murtaza, F.; HaroonYousaf, M.; Velastin, S.A. DA-VLAD: Discriminative Action Vector of Locally Aggregated Descriptors for Action Recognition. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 3993–3997.
42. Rahmani, H.; Mian, A.; Shah, M. Learning a deep model for human action recognition from novel viewpoints. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 667–681. [[CrossRef](#)] [[PubMed](#)]
43. Wang, L.; Xu, Y.; Cheng, J.; Xia, H.; Yin, J.; Wu, J. Human Action Recognition by Learning Spatio-Temporal Features With Deep Neural Networks. *IEEE Access* **2018**, *6*, 17913–17922. [[CrossRef](#)]
44. Khan, M.A.; Akram, T.; Sharif, M.; Javed, M.Y.; Muhammad, N.; Yasmin, M. An implementation of optimized framework for action classification using multilayers neural network on selected fused features. In *Pattern Analysis and Applications*; Springer: London, UK, February 2018; pp. 1–21.
45. Hara, K.; Kataoka, H.; Satoh, Y. Learning spatio-temporal features with 3D residual networks for action recognition. In Proceedings of the ICCV Workshop on Action, Gesture, and Emotion Recognition, Venice, Italy, 29 October 2017; p. 4.

46. Yue-Hei Ng, J.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; Toderici, G. Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–10 June 2015; pp. 4694–4702.
47. Feichtenhofer, C.; Pinz, A.; Wildes, R. Spatiotemporal residual networks for video action recognition. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA 2016; pp. 3468–3476.
48. Feichtenhofer, C.; Pinz, A.; Wildes, R.P. Temporal residual networks for dynamic scene recognition. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 7435–7444.
49. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [[CrossRef](#)]
50. Sun, X.; Chen, M.; Hauptmann, A. Action recognition via local descriptors and holistic features. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Miami, FL, USA, 20–25 June 2009; pp. 58–65.
51. Wagstaff, K.; Cardie, C.; Rogers, S.; Schrödl, S. Constrained k-means clustering with background knowledge. In Proceedings of the Eighteenth International Conference on Machine Learning, Williamstown, MA, USA, 28 June–1 July, 2001; pp. 577–584.
52. Chatfield, K.; Lempitsky, V.S.; Vedaldi, A.; Zisserman, A. The devil is in the details: An evaluation of recent feature encoding methods. In Proceedings of the British Machine Vision Conference, Dundee, UK, 29 August–2 September 2011; p. 8.
53. Escalera, S.; Pujol, O.; Radeva, P. Separability of ternary codes for sparse designs of error-correcting output codes. *Pattern Recognit. Lett.* **2009**, *30*, 285–297. [[CrossRef](#)]
54. Chapelle, O.; Vapnik, V.; Bousquet, O.; Mukherjee, S. Choosing multiple parameters for support vector machines. *Mach. Learn.* **2002**, *46*, 131–159. [[CrossRef](#)]
55. Yi, Y.; Zheng, Z.; Lin, M. Realistic action recognition with salient foreground trajectories. *Expert Syst. Appl.* **2017**, *75*, 44–55. [[CrossRef](#)]
56. Wang, L.; Li, R.; Fang, Y. Power difference template for action recognition. *Mach. Vis. Appl.* **2017**, *28*, 1–11. [[CrossRef](#)]
57. Sheng, B.; Yang, W.; Sun, C. Action recognition using direction-dependent feature pairs and non-negative low rank sparse model. *Neurocomputing* **2015**, *158*, 73–80. [[CrossRef](#)]
58. Ballas, N.; Yang, Y.; Lan, Z.Z.; Delezoide, B.; Prêteux, F.; Hauptmann, A. Space-time robust representation for action recognition. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 2704–2711.
59. Yao, T.; Wang, Z.; Xie, Z.; Gao, J.; Feng, D.D. Learning universal multiview dictionary for human action recognition. *Pattern Recognit.* **2017**, *64*, 236–244. [[CrossRef](#)]
60. Ullah, J.; Jaffar, M.A. Object and motion cues based collaborative approach for human activity localization and recognition in unconstrained videos. *Clust. Comput.* **2017**, *21*, 1–12. [[CrossRef](#)]
61. Tong, M.; Wang, H.; Tian, W.; Yang, S. Action recognition new framework with robust 3D-TCCHOGAC and 3D-HOOFAC. *Multimed. Tools Appl.* **2017**, *76*, 3011–3030. [[CrossRef](#)]
62. Hsieh, C.Y.; Lin, W.Y. Video-based human action and hand gesture recognition by fusing factored matrices of dual tensors. *Multimed. Tools Appl.* **2017**, *76*, 7575–7594. [[CrossRef](#)]
63. Peng, X.; Schmid, C. Multi-region two-stream R-CNN for action detection. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 744–759.
64. Wang, T.; Wang, S.; Ding, X. Detecting human action as the spatio-temporal tube of maximum mutual information. *IEEE Trans. Circuits Syst. Video Technol.* **2014**, *24*, 277–290. [[CrossRef](#)]
65. Cho, J.; Lee, M.; Chang, H.J.; Oh, S. Robust action recognition using local motion and group sparsity. *Pattern Recognit.* **2014**, *47*, 1813–1825. [[CrossRef](#)]
66. Wang, H.; Kläser, A.; Schmid, C.; Liu, C.L. Dense trajectories and motion boundary descriptors for action recognition. *Int. J. Comput. Vis.* **2013**, *103*, 60–79. [[CrossRef](#)]
67. Wang, H.; Yuan, C.; Hu, W.; Sun, C. Supervised class-specific dictionary learning for sparse modeling in action recognition. *Pattern Recognit.* **2012**, *45*, 3902–3911. [[CrossRef](#)]
68. Le, Q.V.; Zou, W.Y.; Yeung, S.Y.; Ng, A.Y. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 20–25 June 2011; pp. 3361–3368.

69. Kläser, A.; Marszałek, M.; Laptev, I.; Schmid, C. *Will Person Detection Help Bag-of-Features Action Recognition?*; INRIA: Rocquencourt, France, 2010.
70. Yi, Y.; Lin, M. Human action recognition with graph-based multiple-instance learning. *Pattern Recognit.* **2016**, *53*, 148–162. [[CrossRef](#)]
71. Duta, I.C.; Uijlings, J.R.; Ionescu, B.; Aizawa, K.; Hauptmann, A.G.; Sebe, N. Efficient human action recognition using histograms of motion gradients and VLAD with descriptor shape information. *Multimed. Tools Appl.* **2017**, *76*, 22445–22472. [[CrossRef](#)]
72. Peng, X.; Wang, L.; Wang, X.; Qiao, Y. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *Comput. Vis. Image Underst.* **2016**, *150*, 109–125. [[CrossRef](#)]



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).